

Chernoff information of exponential families

Frank Nielsen, *Senior Member, IEEE*,

Abstract—Chernoff information upper bounds the probability of error of the optimal Bayesian decision rule for 2-class classification problems. However, it turns out that in practice the Chernoff bound is hard to calculate or even approximate. In statistics, many usual distributions, such as Gaussians, Poissons or frequency histograms called multinomials, can be handled in the unified framework of exponential families. In this note, we prove that the Chernoff information for members of the same exponential family can be either derived analytically in closed form, or efficiently approximated using a simple geodesic bisection optimization technique based on an exact geometric characterization of the “Chernoff point” on the underlying statistical manifold.

Index Terms—Chernoff information, α -divergences, exponential families, information geometry.



1 INTRODUCTION

CONSIDER the following statistical decision problem of *classifying* a random observation x as one of two possible classes: C_1 and C_2 (say, detect target signal from noise signal). Let $w_1 = \Pr(C_1) > 0$ and $w_2 = \Pr(C_2) = 1 - w_1 > 0$ denote the *a priori* class probabilities, and let $p_1(x) = \Pr(x|C_1)$ and $p_2(x) = \Pr(x|C_2)$ denote the *class-conditional* probabilities, so that we have $p(x) = w_1 p_1(x) + w_2 p_2(x)$. Bayes decision rule classifies x as C_1 if $\Pr(C_1|x) > \Pr(C_2|x)$, and as C_2 otherwise. Using Bayes rule¹, we have $\Pr(C_i|x) = \frac{\Pr(C_i)\Pr(x|C_i)}{\Pr(x)} = \frac{w_i p_i(x)}{p(x)}$ for $i \in \{1, 2\}$. Thus Bayes decision rule assigns x to class C_1 if and only if $w_1 p_1(x) > w_2 p_2(x)$, and to C_2 otherwise. Let $L(x) = \frac{\Pr(x|C_1)}{\Pr(x|C_2)}$ denote the *likelihood ratio*. In decision theory [1], Neyman and Pearson proved that the optimum decision *test* has necessarily to be of the form $L(x) \geq t$ to accept hypothesis C_1 , where t is a threshold value.

The probability of error $E = \Pr(\text{Error})$ of *any* decision rule \mathfrak{D} is $E = \int p(x) \Pr(\text{Error}|x) dx$, where

$$\Pr(\text{Error}|x) = \begin{cases} \Pr(C_1|x) & \text{if } \mathfrak{D} \text{ wrongly decided } C_2, \\ \Pr(C_2|x) & \text{if } \mathfrak{D} \text{ wrongly decided } C_1. \end{cases}$$

Thus Bayes decision rule minimizes *by principle* the average *probability of error*:

$$E^* = \int \Pr(\text{Error}|x) p(x) dx, \quad (1)$$

$$= \int \min(\Pr(C_1|x), \Pr(C_2|x)) p(x) dx. \quad (2)$$

The Bayesian rule is also called the maximum a-posteriori (MAP) decision rule. Bayes error constitutes

therefore the reference benchmark since no other decision rule can beat its classification performance.

Bounding tightly the Bayes error is thus crucial in hypothesis testing. Chernoff derived a notion of information² from this hypothesis task (see Section 7 of [2]). To upper bound Bayes error, one replaces the minimum function by a smooth power function: Namely, for $a, b > 0$, we have

$$\min(a, b) \leq a^\alpha b^{1-\alpha}, \forall \alpha \in (0, 1). \quad (3)$$

Thus we get the following Chernoff bound:

$$E^* = \int \min(\Pr(C_1|x), \Pr(C_2|x)) p(x) dx \quad (4)$$

$$\leq w_1^\alpha w_2^{1-\alpha} \int p_1^\alpha(x) p_2^{1-\alpha}(x) dx \quad (5)$$

Since the inequality holds for any $\alpha \in (0, 1)$, we upper bound the minimum error E^* as follows

$$E^* \leq w_1^\alpha w_2^{1-\alpha} c_\alpha(p_1 : p_2),$$

where $c_\alpha(p_1 : p_2) = \int p_1^\alpha(x) p_2^{1-\alpha}(x) dx$ is called the Chernoff α -coefficient. We use the “:” delimiter to emphasize the fact that this statistical measure is usually not symmetric: $c_\alpha(p_1 : p_2) \neq c_\alpha(p_2 : p_1)$, although we have $c_\alpha(p_2 : p_1) = c_{1-\alpha}(p_1 : p_2)$. For $\alpha = \frac{1}{2}$, we obtain the symmetric Bhattacharyya coefficient [3] $b(p_1 : p_2) = c_{\frac{1}{2}}(p_1 : p_2) = \int \sqrt{p_1(x)p_2(x)} dx = b(p_2, p_1)$. The optimal Chernoff α -coefficient is found by choosing the *best* exponent for upper bounding Bayes error [1]:

2. In information theory, there exists several notions of information such as Fisher information in Statistics or Shannon information in Coding theory. Those various definitions gained momentum by asking questions like “How hard is it to estimate/discriminate distributions?” (Fisher) or “How hard is it to compress data?” (Shannon). Those “how hard...” questions were answered by proving lower bounds (Cramér-Rao for Fisher, and Entropy for Shannon). Similarly, Chernoff information answers the “How hard is it to classify (empirical) data?” by providing a tight lower bound: the (Chernoff) (classification) information.

• F. Nielsen is with the Sony Computer Science Laboratories (Tokyo, Japan) and École Polytechnique (Palaiseau, France).
E-mail: nielsen@lix.polytechnique.fr

1. Bayes rule states that the joint probability of two events equals the product of the probability of one event times the conditional probability of the second event given the first one. That is, in mathematical terms $\Pr(x \wedge \theta) = \Pr(x)\Pr(\theta|x) = \Pr(\theta)\Pr(x|\theta)$, so that we have $\Pr(\theta|x) = \Pr(\theta)\Pr(x|\theta)/\Pr(x)$.

$$c^*(p_1 : p_2) = c_{\alpha^*}(p_1 : p_2) = \min_{\alpha \in (0,1)} \int p_1^\alpha(x) p_2^{1-\alpha}(x) dx. \quad (6)$$

Since the Chernoff coefficient is a *measure of similarity* (with $0 < c_\alpha(p_1, p_2) \leq 1$) relating to the overlapping of the densities p_1 and p_2 , it follows that we can derive thereof a statistical distance measure, called the *Chernoff information* (or Chernoff divergence) as

$$C^*(p_1 : p_2) = C_{\alpha^*}(p_1 : p_2) \quad (7)$$

$$= -\log \min_{\alpha \in (0,1)} \int p_1^\alpha(x) p_2^{1-\alpha}(x) dx \geq 0.$$

$$= \max_{\alpha \in (0,1)} -\log \int p_1^\alpha(x) p_2^{1-\alpha}(x) dx \quad (8)$$

In the remainder, we call Chernoff divergence (or Chernoff information) the measure $C^*(\cdot : \cdot)$, and Chernoff α -divergence (of the first type) the functional $C_\alpha(p : q)$ (for $\alpha \in (0,1)$). Chernoff information yields the best achievable exponent for a Bayesian probability of error [1]:

$$E^* \leq w_1^{\alpha^*} w_2^{1-\alpha^*} e^{-C^*(p_1:p_2)}. \quad (9)$$

From the Chernoff α -coefficient measure of similarity, we can derive a second type of Chernoff α -divergences [4] defined by $C'_\alpha(p : q) = \frac{1}{\alpha(1-\alpha)}(1 - c_\alpha(p : q))$. Those second type Chernoff α -divergences are related to Amari α -divergences [5] by a linear mapping [4] on the exponent α , and to Rényi and Tsallis relative entropies (see Section 4). In the remainder, Chernoff α -divergences refer to the first-type divergence.

In practice, we do *not* have statistical knowledge of the prior distributions of classes nor of the class-conditional distributions. But we are rather given a training set of correctly labeled class points. In that case, a simple decision rule, called the *nearest neighbor rule*³, consists for an observation x , to label it according to the label of its nearest neighbor (ground-truth). It can be shown that the probability error of this simple scheme is upper bounded by *twice* the optimal Bayes error [6], [7]. Thus half of the Chernoff information is contained somehow in the nearest neighbor knowledge, a key component of machine learning algorithms. (It is traditional to improve this classification by taking a majority vote over the k nearest neighbors.)

Chernoff information has appeared in many applications ranging from sensor networks [8] to visual computing tasks such as image segmentation [9], image registration [10], face recognition [11], feature detector [12], and edge segmentation [13], just to name a few.

The paper is organized as follows: Section 2 introduces the functional parametric Bregman and Jensen class of statistical distances. Section 3 concisely describes the exponential families in statistics. Section 4 proves that

the Chernoff α -divergences of two members of the *same* exponential family class is equivalent to a skew Jensen divergence evaluated at the corresponding distribution parameters. In section 5, we show that the optimal Chernoff coefficient obtained by minimizing skew Jensen divergences yields an equivalent Bregman divergence, which can be derived from a simple optimality criterion. It follows a closed-form formula for the Chernoff information on single-parametric exponential families in Section 5.1. We extend the optimality criterion to the multi-parametric case in Section 5.2. Section 6 characterizes geometrically the optimal solution by introducing concepts of information geometry. Section 7 designs a simple yet efficient geodesic bisection search algorithm for approximating the multi-parametric case. Finally, section 8 concludes the paper.

2 STATISTICAL DIVERGENCES

Given two probability distributions with respective densities p and q , a divergence $D(p : q)$ measures the distance between those distributions. The classical divergence in information theory [1] is the *Kullback-Leibler divergence*, also called *relative entropy*:

$$\text{KL}(p : q) = \int p(x) \log \frac{p(x)}{q(x)} dx \quad (10)$$

(For probability mass functions, the integral is replaced by a discrete sum.) This divergence is oriented (ie. $\text{KL}(p : q) \neq \text{KL}(q : p)$) and does not satisfy the triangle inequality of metrics. It turns out that the Kullback-Leibler divergence belongs to a wider class of divergences called Bregman divergences. A Bregman divergence is obtained for a strictly convex and differentiable generator F as:

$$B_F(p : q) = \int (F(p(x)) - F(q(x)) - (p(x) - q(x))F'(q(x))) dx \quad (11)$$

The Kullback-Leibler divergence is obtained for the generator $F(x) = x \log x$, the negative Shannon entropy (also called Shannon information). This functional parametric class of Bregman divergences B_F can further be interpreted as *limit cases* of skew Jensen divergences. A skew Jensen divergence (Jensen α -divergences, or α -Jensen divergences) is defined for a strictly convex generator F as

$$J_F^{(\alpha)}(p : q) = \int (\alpha F(p(x)) + (1 - \alpha)F(q(x)) - F(\alpha p(x) + (1 - \alpha)q(x))) dx \geq 0, \quad \forall \alpha \in (0,1) \quad (12)$$

Note that $J_F^{(\alpha)}(p : q) = J_F^{(1-\alpha)}(q : p)$, and that F is defined up to affine terms. For $\alpha \rightarrow \{0,1\}$, the Jensen divergence tend to zero, and loose its power of discrimination. However, interestingly, we have $\lim_{\alpha \rightarrow 1} J_F^{(\alpha)}(p : q) = \frac{1}{1-\alpha} B_F(p : q)$ and $\lim_{\alpha \rightarrow 0} J_F^{(\alpha)}(p : q) = \frac{1}{\alpha} B_F(q : p)$,

3. The nearest neighbor rule postulates that things that “look alike must be alike.” See [6].

as proved in [14], [15]. That is, Jensen α -divergences tend asymptotically to (scaled) Bregman divergences.

The Kullback-Leibler divergence also belongs to the class of Csiszár F -divergences (with $F(x) = x \log x$), defined for a convex function F with $F(1) = 0$:

$$I_F(p : q) = \int_x F\left(\frac{p(x)}{q(x)}\right) q(x) dx. \quad (13)$$

Amari's α -divergences are the canonical divergences in α -flat spaces in information geometry [16] defined by

$$A_\alpha(p : q) = \begin{cases} \frac{4}{1-\alpha^2}(1 - c_{1-\alpha}(p : q)), & \alpha \neq \pm 1, \\ \int p(x) \log \frac{p(x)}{q(x)} dx = \text{KL}(p, q), & \alpha = -1, \\ \int q(x) \log \frac{q(x)}{p(x)} dx = \text{KL}(q, p), & \alpha = 1, \end{cases} \quad (14)$$

Those Amari α -divergences (related to Chernoff α -coefficients, and Chernoff α -divergences of the second type by a linear mapping of the exponent [4]) are F -divergences for the generator $F_\alpha(x) = \frac{4}{1-\alpha^2}(1 - x^{\frac{1+\alpha}{2}})$, $\alpha \notin \{-1, 1\}$.

Next, we introduce a versatile class of probability densities in statistics for which α -Jensen divergences (and hence Bregman divergences) admit closed-form formula.

3 EXPONENTIAL FAMILIES

A generic class of statistical distributions encapsulating many usual distributions (Bernoulli, Poisson, Gaussian, multinomials, Beta, Gamma, Dirichlet, etc.) are the exponential families. We recall their elementary definition here, and refer the reader to [17] for a more detailed overview. An *exponential family* E_F is a parametric set of probability distributions admitting the following canonical decomposition of their densities:

$$p(x; \theta) = \exp(\langle t(x), \theta \rangle - F(\theta) + k(x)) \quad (15)$$

where $t(x)$ is the sufficient statistic, $\theta \in \Theta$ are the natural parameters belonging to an open convex natural space Θ , $\langle \cdot, \cdot \rangle$ is the inner product (i.e., $\langle x, y \rangle = x^T y$ for column vectors), $F(\cdot)$ is the log-normalizer (a C^∞ convex function), and $k(x)$ the carrier measure.

For example, Poisson distributions $\Pr(x = k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$, for $k \in \mathbb{N}$ form an exponential family $E_F = \{p_F(x; \theta) \mid \theta \in \Theta\}$, with $t(x) = x$ the sufficient statistic, $\theta = \log \lambda$ the natural parameters, $F(\theta) = \exp \theta$ the log-normalizer, and $k(x) = -\log x!$ is the carrier measure.

Since we often deal with applications using multivariate normals, we also report the canonical decomposition for the multivariate Gaussian family. We rewrite the Gaussian density of mean μ and variance-covariance matrix Σ :

$$p(x; \mu, \Sigma) = \frac{1}{2\pi\sqrt{\det \Sigma}} \exp\left(-\frac{(x - \mu)^T \Sigma^{-1} (x - \mu)}{2}\right)$$

in the canonical form with $\theta = (\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1}) \in \Theta = \mathbb{R}^d \times \mathbb{K}_{d \times d}$ ($\mathbb{K}_{d \times d}$ denotes the cone of positive definite matrices), $F(\theta) = \frac{1}{4}\text{tr}(\theta_2^{-1}\theta_1\theta_1^T) - \frac{1}{2}\log \det \theta_2 + \frac{d}{2}\log \pi$ the

log-normalizer, $t(x) = (x, -x^T x)$ the sufficient statistics, and $k(x) = 0$ the carrier measure. In that case, the inner product $\langle \cdot, \cdot \rangle$ is composite, and calculated as the sum of a vector dot product with a matrix trace product: $\langle \theta, \theta' \rangle = \theta_1^T \theta'_1 + \text{tr}(\theta_2^T \theta'_2)$, where $\theta = [\theta_1 \ \theta_2]^T$ and $\theta' = [\theta'_1 \ \theta'_2]^T$.

The *order* of an exponential family denotes the dimension of its parameter space. For example, Poisson family is of order 1, univariate Gaussians of order 2, and d -dimensional multivariate Gaussians of order $\frac{d(d+3)}{2}$. Exponential families brings mathematical convenience to easily solve tasks, like finding the maximum likelihood estimators [17]. It can be shown that the Kullback-Leibler divergence of members of the same exponential family is equivalent to a Bregman divergence on the natural parameters [18], thus bypassing the fastidious integral computation of Eq. 10, and yielding a closed-form formula (following Eq. 11):

$$\text{KL}(p_F(x; \theta_p) : p_F(x; \theta_q)) = B_F(\theta_q : \theta_p). \quad (16)$$

Note that on the left hand side, the Kullback-Leibler is a distance acting on distributions, while on the right hand side, the Bregman divergence is a distance acting on corresponding swapper parameters.

Exponential families play a crucial role in statistics as they also bring mathematical convenience for generalizing results. For example, the log-likelihood ratio test for members of the same exponential family writes down as:

$$\log \frac{e^{\langle t(x), \theta_1 \rangle - F(\theta_1) + k(x)}}{e^{\langle t(x), \theta_2 \rangle - F(\theta_2) + k(x)}} \geq \log \frac{w_2}{w_1} \quad (17)$$

Thus the decision border is a *linear bisector* in the sufficient statistics $t(x)$:

$$\langle t(x), \theta_1 - \theta_2 \rangle - F(\theta_1) + F(\theta_2) = \log \frac{w_2}{w_1}. \quad (18)$$

4 CHERNOFF COEFFICIENTS OF EXPONENTIAL FAMILIES

Let us prove that the Chernoff α -divergence of members of the *same* exponential families is equivalent to a α -Jensen divergence defined for the log-normalizer generator, and evaluated at the corresponding natural parameters. Without loss of generality, let us consider the reduced canonical form of exponential families $p_F(x; \theta) = \exp(\langle x, \theta \rangle - F(\theta))$ (assuming $t(x) = x$ and $k(x) = 0$). Consider the Chernoff α -coefficient of similarity of two distributions p and q belonging to the *same* exponential family E_F :

$$c_\alpha(p : q) = \int p^\alpha(x) q^{1-\alpha}(x) dx = \int p_F^{(\alpha)}(x; \theta_p) p_F^{1-\alpha}(x; \theta_q) dx \quad (19)$$

$$\begin{aligned}
&= \int \exp(\alpha(\langle x, \theta_p \rangle - F(\theta_p))) \exp((1-\alpha)(\langle x, \theta_q \rangle - F(\theta_q))) dx \quad (\text{Note that } R_{\frac{1}{2}}(p : q) \text{ is twice the Bhattacharyya coefficient: } R_{\frac{1}{2}}(p : q) = 2C_{\frac{1}{2}}(p : q).) \text{ For example, the Rényi divergence on members } p \sim N(\mu_p, \Sigma_p) \text{ and } q \sim N(\mu_q, \Sigma_q) \text{ of the normal exponential family is obtained in closed form solution using Eq. 24:} \\
&= \int \exp(\langle x, \alpha\theta_p + (1-\alpha)\theta_q \rangle - (\alpha F(\theta_p) + (1-\alpha)F(\theta_q))) dx \\
&= \exp(-(\alpha F(\theta_p) + (1-\alpha)F(\theta_q))) \int \exp(\langle x, \alpha\theta_p + (1-\alpha)\theta_q \rangle - F(\alpha\theta_p + (1-\alpha)\theta_q)) dx \\
&= \exp(F(\alpha\theta_p + (1-\alpha)\theta_q) - (\alpha F(\theta_p) + (1-\alpha)F(\theta_q))) \times \\
&\quad \int \exp(\langle x, \alpha\theta_p + (1-\alpha)\theta_q \rangle - F(\alpha\theta_p + (1-\alpha)\theta_q)) dx \\
&= \exp(F(\alpha\theta_p + (1-\alpha)\theta_q) - (\alpha F(\theta_p) + (1-\alpha)F(\theta_q))) \times \int p_F(x; \alpha\theta_p + (1-\alpha)\theta_q) dx \\
&= \exp(-J_F^{(\alpha)}(\theta_p : \theta_q)) \geq 0.
\end{aligned}$$

It follows that the Chernoff α -divergence (of the first type) is given by

$$\begin{aligned}
C_\alpha(p : q) &= -\log c_\alpha(p, q) = J_F^{(\alpha)}(\theta_p : \theta_q), \\
c_\alpha(p : q) &= e^{-C_\alpha(p : q)} = e^{-J_F^{(\alpha)}(\theta_p : \theta_q)}.
\end{aligned}$$

That is, the Chernoff α -divergence on members of the same exponential family is equivalent to a Jensen α -divergence on the corresponding natural parameters. For multivariate normals, we thus retrieve easily the following Chernoff α -divergence between $p \sim N(\mu_1, \Sigma_1)$ and $q \sim N(\mu_2, \Sigma_2)$:

$$\begin{aligned}
C_\alpha(p, q) &= \frac{1}{2} \log \frac{|\alpha\Sigma_1 + (1-\alpha)\Sigma_2|}{|\Sigma_1|^\alpha |\Sigma_2|^{1-\alpha}} + \\
&\quad \frac{\alpha(1-\alpha)}{2} (\mu_1 - \mu_2)^T (\alpha\Sigma_1 + (1-\alpha)\Sigma_2) (\mu_1 - \mu_2).
\end{aligned} \quad (20)$$

For $\alpha = \frac{1}{2}$, we find the Bhattacharyya distance [3], [19] between multivariate Gaussians.

Note that since Chernoff α -divergences are related to Rényi α -divergences

$$R_\alpha(p : q) = \frac{1}{\alpha - 1} \log \int_x p(x)^\alpha q^{1-\alpha}(x) dx, \quad (21)$$

built on Rényi entropy

$$H_R^\alpha(p) = \frac{1}{1-\alpha} \log \left(\int_x p^\alpha(x) dx - 1 \right), \quad (22)$$

(and hence by a monotonic mapping⁴ to Tsallis divergences), closed form formulas for members of the same exponential family follow:

$$R_\alpha(p : q) = \frac{1}{1-\alpha} C_\alpha(p : q), \quad (23)$$

$$R_\alpha(p_F(x; \theta_p) : p_F(x; \theta_q)) = \frac{1}{1-\alpha} J_F^{(\alpha)}(\theta_p : \theta_q) \quad (24)$$

4. The Tsallis entropy $H_T^\alpha(p) = \frac{1}{\alpha-1} (1 - \int p(x)^\alpha dx)$ is obtained from the Rényi entropy (and vice-versa) via the mappings: $H_T^\alpha(p) = \frac{1}{1-\alpha} (e^{(1-\alpha)H_R^\alpha(p)} - 1)$ and $H_R^\alpha(p) = \frac{1}{1-\alpha} \log(1 + (1-\alpha)H_T^\alpha(p))$.

$$R_\alpha(p, q) = \frac{1}{2\alpha} (\mu_p - \mu_q)^T ((1-\alpha)\Sigma_p + \alpha\Sigma_q)^{-1} (\mu_p - \mu_q) + \frac{1}{1-\alpha} \log \frac{\det((1-\alpha)\Sigma_p + \alpha\Sigma_q)}{\det(\Sigma_p^{1-\alpha}) \det(\Sigma_q^\alpha)}. \quad (25)$$

Similarly, for the Tsallis relative entropy, we have:

$$\begin{aligned}
T_\alpha(p : q) &= \frac{1}{1-\alpha} (1 - c_\alpha(q : p)), \quad (26) \\
T_\alpha(p_F(x; \theta_p) : p_F(x; \theta_q)) &= \frac{(1 - e^{-J_F^{(\alpha)}(q:p)})}{1-\alpha} \quad (27)
\end{aligned}$$

Note that $\lim_{\alpha \rightarrow 1} R_\alpha(p : q) = \lim_{\alpha \rightarrow 1} T_\alpha(p : q) = \text{KL}(p : q) = B_F(\theta_q : \theta_p)$, as expected.

So far, particular cases of exponential families have been considered for computing the Chernoff α -divergences (but not Chernoff divergence). For example, Rauber et al. [20] investigated statistical distances for Dirichlet and Beta distributions (both belonging to the exponential families). The density of a Dirichlet distribution parameterized by a d -dimensional vector $p = (p_1, \dots, p_d)$ is

$$\Pr(X = x; p) = \frac{\Gamma(\sum_{i=1}^d p_i)}{\prod_{i=1}^d \Gamma(p_i)} \prod_{i=1}^d x_i^{p_i-1},$$

with $\Gamma(t) = \int_0^\infty z^{t-1} e^{-z} dz$ the gamma function generalizing the factorial $\Gamma(n-1) = n!$. Beta distributions are particular cases of Dirichlet distributions, obtained for $d = 2$. Rauber et al. [20] report the following closed-form formula for the Chernoff α -divergences:

$$\begin{aligned}
C_\alpha(p : q) &= \log \Gamma \left(\sum_{i=1}^d (\alpha p_i - (1-\alpha) q_i) \right) \\
&+ \alpha \sum_{i=1}^d \log \Gamma(p_i) + (1-\alpha) \sum_{i=1}^d \log \Gamma(q_i) \\
&- \sum_{i=1}^d \log \Gamma(\alpha p_i - (1-\alpha) q_i) - \\
&\alpha \log \Gamma \left(\sum_{i=1}^d |p_i| \right) - (1-\alpha) \log \Gamma \left(\sum_{i=1}^d |q_i| \right).
\end{aligned}$$

Dirichlet distributions are exponential families of order d with natural parameters $\theta = (p_1 - 1, \dots, p_d - 1)$ and log-normalizer $F(\theta) = \sum_{i=1}^d \log \Gamma(\theta_i + 1) - \log \Gamma(d + \sum_{i=1}^d \theta_i)$ (or $F(p) = \sum_{i=1}^d \log \Gamma(p_i) - \log \Gamma(\sum_{i=1}^d p_i)$). Our work extends the computation of Chernoff α -divergences to *arbitrary* exponential families using the natural parameters and the log-normalizer.

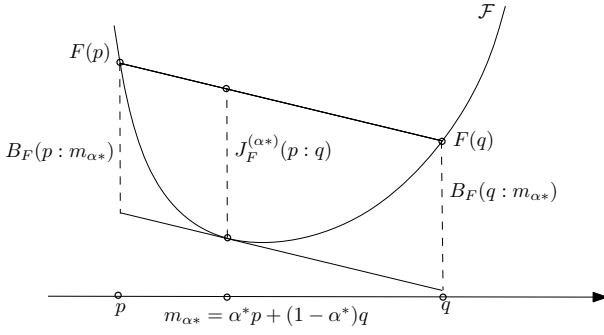


Fig. 1. The maximal Jensen α -divergence is a Bregman divergence in disguise: $J_F^{(\alpha^*)}(p:q) = \max_{\alpha \in (0,1)} J_F^{(\alpha)}(p:q) = B_F(p:m_{\alpha^*}) = B_F(q:m_{\alpha^*})$.

Since Chernoff information is defined as the *maximal* Chernoff α -divergence (which corresponds to minimize the Chernoff coefficient in the Bayes error upper bound, with $0 < c_\alpha(p,q) \leq 1$), we concentrate on maximizing the equivalent skew Jensen divergence.

5 MAXIMIZING α -JENSEN DIVERGENCES

We now prove that the maximal skew Jensen divergence can be computed as an *equivalent* Bregman divergence. First, consider univariate functions. Let $\alpha^* = \arg \max_{0 < \alpha < 1} J_F^{(\alpha)}(p:q)$ be the maximal α -divergence. Following Figure 1, we observe that we have *geometrically* the following relationships [14]:

$$J_F^{(\alpha^*)}(p:q) = B_F(p:m_{\alpha^*}) = B_F(q:m_{\alpha^*}), \quad (28)$$

where $m_\alpha = \alpha p + (1-\alpha)q$ be the α -mixing of distributions p and q . We maximize the α -Jensen divergence by setting its derivative to zero:

$$\frac{dJ_F^{(\alpha)}(p:q)}{d\alpha} = F(p) - F(q) - (m_\alpha)' F'(m_\alpha). \quad (29)$$

Since the derivative $(m_\alpha)'$ of m_α is equal to $p - q$, we deduce from the maximization that $\frac{dJ_F^{(\alpha)}(p:q)}{d\alpha} = 0$ implies the following constraint:

$$F'(m_\alpha^*) = \frac{F(p) - F(q)}{p - q}. \quad (30)$$

This means geometrically that the tangent at α^* should be parallel to the line passing through $(p, z = F(p))$ and $(q, z = F(q))$, as illustrated in Figure 1. It follows that

$$\alpha^* = \frac{F'^{-1}\left(\frac{F(p)-F(q)}{p-q}\right) - p}{q - p}. \quad (31)$$

Using Eq. 28, we have $p - m_\alpha^* = (1 - \alpha^*)(p - q)$, so that it comes

$$\begin{aligned} B_F(p:m_\alpha^*) &= F(p) - F(m_\alpha^*) - (1 - \alpha^*)(F(p) - F(q)) \\ &= \alpha^* F(p) + (1 - \alpha^*) F(q) - F(m_\alpha^*) \\ &= J_F^{(\alpha^*)}(p:q) \end{aligned}$$

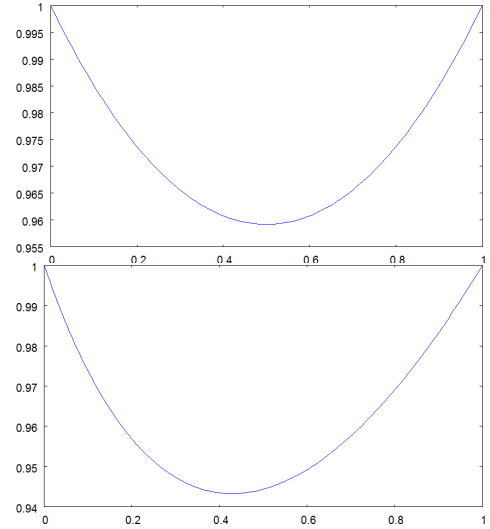


Fig. 2. Plot of the α -divergences for two normal distributions for $\alpha \in (0, 1)$: (Top) $p \sim N(0, 9)$ and $q \sim N(2, 9)$, and (Bottom) $p \sim N(0, 9)$ and $q \sim N(2, 36)$. Observe that for equal variance, the minimum α divergence is obtained for $\alpha = \frac{1}{2}$, and that Chernoff divergence reduces to the Bhattacharyya divergence.

Similarly, we have $q - m_\alpha^* = \alpha^*(q - p)$ and it follows that

$$\begin{aligned} B_F(q:m_\alpha^*) &= F(q) - F(m_\alpha^*) - (q - m_\alpha^*) F'(m_\alpha^*) \\ &= F(q) - F(m_\alpha^*) + \alpha^*(p - q) \frac{F(p) - F(q)}{p - q} \\ &= \alpha^* F(p) + (1 - \alpha^*) F(q) - F(m_\alpha^*) \\ &= J_F^{(\alpha^*)}(p:q) \end{aligned} \quad (32)$$

Thus, we *analytically* checked the geometric intuition that $J_F^{(\alpha^*)}(p:q) = B_F(p:m_\alpha^*) = B_F(q:m_\alpha^*)$. Observe that in the definition of a Bregman divergence, we require to compute explicitly the gradient ∇F , but that in the Jensen α -divergence, we do not need it. (However, the gradient computation occurs in the computation of the best α).

5.1 Single-parametric exponential families

We conclude that the Chernoff information divergence of members of the same exponential family of order 1 has always a closed-form analytic formula:

$$C(p:q) = \alpha^* F(p) + (1 - \alpha^*) F(q) - F\left(F'^{-1}\left(\frac{F(p) - F(q)}{p - q}\right)\right), \quad (33)$$

with

$$\alpha^* = \frac{F'^{-1}\left(\frac{F(p)-F(q)}{p-q}\right) - p}{q - p}. \quad (34)$$

Common exponential families of order 1 include the Binomial, Bernoulli, Laplacian (exponential), Rayleigh,

Poisson, Gaussian with fixed standard deviations. To illustrate the calculation method, let us instantiate the univariate Gaussian and Poisson distributions.

For univariate Gaussian differing in mean only (ie., constant standard deviation σ), we have the following:

$$\theta = \frac{\mu}{\sigma^2}, \quad F(\theta) = \frac{\theta^2 \sigma^2}{2} = \frac{\mu^2}{2\sigma^2}, \quad F'(\theta) = \theta \sigma^2 = \mu$$

We solve for α^* using Eq. 34:

$$\begin{aligned} F'(\alpha^* \theta_p + (1 - \alpha^*) \theta_q) &= \frac{F(\theta_p) - F(\theta_q)}{\theta_p - \theta_q} \\ \mu_p + (1 - \alpha^*)(\mu_q - \mu_p) &= \frac{\mu_p^2 - \mu_q^2}{2(\mu_p - \mu_q)} = \frac{\mu_p + \mu_q}{2} \end{aligned}$$

It follows that $\alpha^* = \frac{1}{2}$ as expected, and that the Chernoff information is the Bhattacharyya distance:

$$\begin{aligned} C(p : q) &= C_{\frac{1}{2}}(p, q) = J_F^{(\frac{1}{2})}(\theta_p : \theta_q), \\ &= \frac{1}{2\sigma^2} \left(\frac{\mu_p^2 + \mu_q^2}{2} \right) - \frac{(\frac{\mu_p + \mu_q}{2})^2}{2\sigma^2} \\ &= \frac{1}{8\sigma^2} (\mu_p - \mu_q)^2 \end{aligned}$$

For Poisson distributions ($F(\theta) = \exp(\theta) = F(\log \lambda) = \exp \log \lambda = \lambda$), Chernoff divergence is found by first computing

$$\alpha^* = \frac{\log \frac{\frac{\lambda_2}{\lambda_1} - 1}{\log \frac{\lambda_2}{\lambda_1}}}{\log \frac{\lambda_2}{\lambda_1}}. \quad (35)$$

Then using Eq. 34, we deduce that

$$\begin{aligned} C(\lambda_1 : \lambda_2) &= \lambda_2 + \alpha^*(\lambda_1 - \lambda_2) - \exp(m_{\alpha^*}) \\ &= \lambda_2 + \alpha^*(\lambda_1 - \lambda_2) - \\ &\quad \exp(\alpha^*(\log \lambda_1) + (1 - \alpha^*) \log \lambda_2) \\ &= \lambda_2 + \alpha^*(\lambda_1 - \lambda_2) - \lambda_1^{\alpha^*} \lambda_2^{1-\alpha^*} \end{aligned} \quad (36)$$

Plugging Eq. 35 in Eq. 36, and “beautifying” the formula yields the following closed-form solution for the Chernoff information:

$$C(\lambda_1, \lambda_2) = \lambda_1 \frac{(\frac{\lambda_2}{\lambda_1} - 1)(\log \frac{\frac{\lambda_2}{\lambda_1} - 1}{\log \frac{\lambda_2}{\lambda_1}} - 1) + \log \frac{\lambda_2}{\lambda_1}}{\log \frac{\lambda_2}{\lambda_1}}. \quad (37)$$

5.2 Arbitrary exponential families

For *multivariate* generators F , we consider the restricted univariate convex function $F_{pq}(\alpha) = F(p + (1 - \alpha)(q - p))$ with parameters $p' = 0$ and $q' = 1$, so that $F_{pq}(0) = F(p)$ and $F_{pq}(1) = F(q)$. We have

$$C_F(p : q) = \max_{\alpha} J_F^{(\alpha)}(\theta_p : \theta_q) = \max_{\alpha} J_{F_{\theta_p \theta_q}}^{(\alpha)}(0 : 1). \quad (38)$$

We have $F'_{pq}(\alpha) = (p - q)^T \nabla F(\alpha p + (1 - \alpha)q)$. To get the inverse of F'_{pq} , we need to solve the equation:

$$(p - q)^T \nabla F(\alpha^* p + (1 - \alpha^*)q) = F(q) - F(p). \quad (39)$$

Observe that in 1D, this equation matches Eq. 30. Finding α^* may not always be in closed-form. Let $\theta^* = \alpha^* p + (1 - \alpha^*)q$, then we need to find α^* such that

$$(p - q)^T \nabla F(\theta^*) = F(q) - F(p). \quad (40)$$

Now, observe that equation 40 is equivalent to the following condition:

$$B_F(\theta_p : \theta^*) = B_F(\theta_q : \theta^*) \quad (41)$$

and that therefore it follows that

$$\text{KL}(p_F(x; \theta^*) : p_F(x; \theta_p)) = \text{KL}(p_F(x; \theta^*) : p_F(x; \theta_q)). \quad (42)$$

Thus it can be checked that the Chernoff distribution $r^* = p_F(x; \theta^*)$ is written as

$$p_F(x; \theta^*) = \frac{p_F(x; \theta_p)^{\alpha^*}(x) p_F(x; \theta_q)^{1-\alpha^*}}{\int_x p_F(x; \theta_p)^{\alpha^*}(x) p_F(x; \theta_q)^{1-\alpha^*} dx} \quad (43)$$

6 THE CHERNOFF POINT

Let us consider now the exponential family

$$E_F = \{p_F(x; \theta) \mid \theta \in \Theta\}, \quad (44)$$

as a smooth statistical manifold [16]. Two distributions $p = p_F(x; \theta_p)$ and $q = p_F(x; \theta_q)$ are geometrically viewed as two points (expressed as θ_p and θ_q coordinates in the natural coordinate system). The Kullback-Leibler divergence between p and q is equivalent to a Bregman divergence on the natural parameters: $\text{KL}(p : q) = B_F(\theta_q : \theta_p)$. For infinitesimal close distributions $p \simeq q$, the Fisher information provides the underlying Riemannian metric, and is equal to the Hessian $\nabla^2 F(\theta)$ of the log-normalizer for exponential families [16]. On statistical manifolds [16], we define *two types* of geodesics: the mixture $\nabla^{(m)}$ geodesic and the exponential $\nabla^{(e)}$ geodesics:

$$\nabla^{(m)}(p(x), q(x), \lambda) = (1 - \lambda)p(x) + \lambda q(x), \quad (45)$$

$$\nabla^{(e)}(p(x), q(x), \lambda) = \frac{p(x)^{1-\lambda} q(x)^{\lambda}}{\int_x p(x)^{1-\lambda} q(x)^{\lambda} dx}, \quad (46)$$

$$(47)$$

Furthermore, to any convex function F , we can associate a dual convex conjugate F^* (such that $F^{**} = F$) via the Legendre-Fenchel transformation:

$$F^*(y) = \max_x \{ \langle x, y \rangle - F(x) \}. \quad (48)$$

The maximum is obtained for $y = \nabla F(x)$. Moreover, the convex conjugates are coupled by reciprocal inverse gradient: $\nabla F^* = (\nabla F)^{-1}$. Thus a member p of the exponential family, can be parameterized by its natural

coordinates $\theta_p = \theta(p)$, or dually by its expectation coordinates $\eta_p = \eta(p) = \nabla F(\theta)$. That is, there exists a *dual coordinate system* on the information manifold E_F of the exponential family.

Note that the Chernoff distribution $r^* = p_F(x; \theta^*)$ of Eq. 43 is a distribution belonging to the exponential geodesic. The natural parameters on the exponential geodesic are interpolated linearly in the θ -coordinate system. Thus the exponential geodesic segment has natural coordinates $\theta(p, q, \lambda) = (1 - \lambda)\theta_p + \lambda\theta_q$. Using the dual expectation parameterization $\eta^* = \nabla F(\theta^*)$, we may also rewrite the optimality criterion of equation Eq. 40 equivalently as

$$(p - q)^T \eta^* = F(q) - F(p), \quad (49)$$

with η^* a point on the exponential geodesic parameterized by the expectation parameters (each mixture/exponential geodesic can be parameterized in each natural/-expectation coordinate systems).

From Eq. 41, we deduce that the Chernoff distribution should also necessarily belong to the right-sided Bregman Voronoi bisector

$$V(p, q) = \{x \mid B_F(\theta_p : \theta_x) = B_F(\theta_q : \theta_x)\}. \quad (50)$$

This bisector is curved in the natural coordinate system, but affine in the dual expectation coordinate system [18]. Moreover, we have $B_F(q : p) = B_{F^*}(\nabla F(p) : \nabla F(q))$, so that we may express the right-sided bisector equivalently in the expectation coordinate system as

$$V(p, q) = \{x \mid B_{F^*}(\eta_x : \eta_p) = B_{F^*}(\eta_x : \eta_q)\}. \quad (51)$$

That is, a left-sided bisector for the dual Legendre convex conjugate F^* .

Thus the Chernoff distribution r^* is viewed as a *Chernoff point* on the statistical manifold such that r^* is defined as the intersection of the exponential geodesic (η -geodesic, or e -geodesic) with the curved bisector $\{x \mid B_F(\theta_p : \theta_x) = B_F(\theta_q : \theta_x)\}$. In [18], it is proved that the exponential geodesic right-sided bisector intersection is Bregman orthogonal. Figure 3 illustrates the geometric property of the Chernoff distribution (which can be viewed indifferently in the natural/expectation parameter space), from which the corresponding best exponent can be retrieved to define the Chernoff information.

We following section builds on this *exact geometric characterization* to build a geodesic bisection optimization method to arbitrarily finely approximate the optimal exponent.

7 A GEODESIC BISECTION ALGORITHM

To find the Chernoff point r^* (ie., the parameter $\theta^* = (1 - \alpha^*)\theta_p + \alpha^*\theta_q$, a simple bisection algorithm follows: Let initially $\alpha \in [\alpha_m, \alpha_M]$ with $\alpha_m = 0, \alpha_M = 1$. Compute the midpoint $\alpha' = \frac{\alpha_m + \alpha_M}{2}$ and let $\theta = \theta_p + \alpha'(\theta_q - \theta_p)$. If $B_F(\theta_p : \theta) < B_F(\theta_q : \theta)$ recurse on interval $[\alpha', \alpha_M]$, otherwise recurse on interval $[\alpha_m, \alpha']$. At each stage we

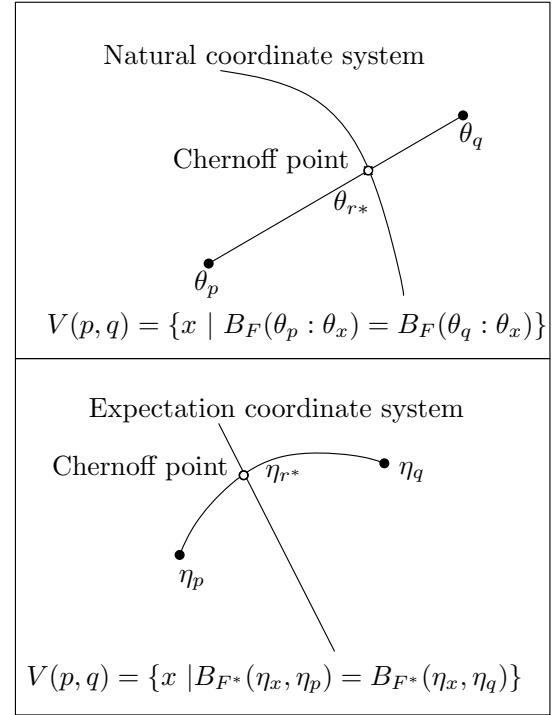


Fig. 3. Chernoff point r^* of p and q is defined as the intersection of the exponential geodesic $\nabla^{(e)}(p, q)$ with the right-sided Voronoi bisector $V(p, q)$. In the natural coordinate system, the exponential geodesic is a line segment and the right-sided bisector is curved. In the dual expectation coordinate system, the exponential geodesic is curved, and the right-sided bisector is affine.

split the α -range in the θ -coordinate system. Thus we can get arbitrarily precise approximation of the Chernoff information of members of the same exponential family by walking on the exponential geodesic towards the Chernoff point.

8 CONCLUDING REMARKS

Chernoff divergence upper bounds asymptotically the optimal Bayes error [1]: $\lim_{n \rightarrow \infty} E^* = e^{-nC(p; q)}$. Chernoff bound thus provides the best Bayesian exponent error [1], improving over the Bhattacharyya divergence ($\alpha = \frac{1}{2}$):

$$\lim_{n \rightarrow \infty} E^* = e^{-nC(p; q)} \leq e^{-nB(p; q)}, \quad (52)$$

at the expense of solving an optimization problem. The probability of misclassification error can also be lower bounded by information-theoretic statistical distances [21], [22] (Stein lemma [1]):

$$\lim_{n \rightarrow \infty} E^* = e^{-nC(p; q)} \geq e^{-nR(p; q)} \geq e^{-nJ(p; q)}, \quad (53)$$

where $J(p : q)$ denotes half of the Jeffreys divergence $J(p : q) = \frac{\text{KL}(p:q) + \text{KL}(q:p)}{2}$ (ie., the arithmetic mean on sided relative entropies) and $R(p : q) = \frac{1}{\frac{1}{\text{KL}(p:q)} + \frac{1}{\text{KL}(q:p)}}$ is the resistor-average distance [22] (ie., the harmonic

mean). In this paper, we have shown that the Chernoff α -divergence of members of the same exponential family can be computed from an equivalent α -Jensen divergence on corresponding natural parameters. Then we have explained how the maximum α -Jensen divergence yields a simple gradient constraint. As a byproduct this shows that the maximal α -Jensen divergence is equivalent to compute a Bregman divergence. For single-parametric exponential families (order-1 families or dimension-wise separable families), we deduced a closed form formula for the Chernoff divergence (or Chernoff information). Otherwise, based on the framework of information geometry, we interpreted the optimization task as of finding the “Chernoff point” defined by the intersection of the exponential geodesic linking the source distributions with a right-sided Bregman Voronoi bisector. Based on this observation, we designed an efficient geodesic bisection algorithm to arbitrarily approximate the Chernoff information.

REFERENCES

- [1] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York, NY, USA: Wiley-Interscience, 1991.
- [2] H. Chernoff, “A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations,” vol. 23, pp. 493–507, 1952.
- [3] A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions,” *Bulletin of Calcutta Mathematical Society*, vol. 35, pp. 99–110, 1943.
- [4] A. Cichocki and S.-i. Amari, “Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities,” *Entropy*, vol. 12, no. 6, pp. 1532–1568, June 2010. [Online]. Available: <http://dx.doi.org/10.3390/e12061532>
- [5] A. Cichocki and S. ichi Amari, “Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities,” *Entropy*, 2010, review submitted.
- [6] T. M. Cover and P. E. Hart, “Nearest neighbor pattern classification,” vol. 13, no. 1, pp. 21–27, 1967.
- [7] C. J. Stone, “Consistent nonparametric regression,” *Annals of Statistics*, vol. 5, no. 4, pp. 595–645, 1977.
- [8] J.-F. Chamberland and V. Veeravalli, “Decentralized detection in sensor networks,” *IEEE Transactions on Signal Processing*, vol. 51(2), pp. 407–416, Feb 2003.
- [9] F. Calderero and F. Marques, “Region merging techniques using information theory statistical measures,” *Transactions on Image Processing*, vol. 19, no. 6, pp. 1567–1586, 2010.
- [10] F. Sadjadi, “Performance evaluations of correlations of digital images using different separability measures,” *IEEE Transactions PAMI*, vol. 4, no. 4, pp. 436–441, Jul. 1982.
- [11] S. K. Zhou and R. Chellappa, “Beyond one still image: Face recognition from multiple still images or video sequence,” in *Face Processing: Advanced Modeling and Methods*. Academic Press, 2005.
- [12] P. Suau and F. Escolano, “Exploiting information theory for filtering the kadir scale-saliency detector,” in *IbPRIA '07: Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part II*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 146–153.
- [13] S. Konishi, A. L. Yuille, J. M. Coughlan, and S. C. Zhu, “Statistical edge detection: Learning and evaluating edge cues,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 57–74, 2003.
- [14] M. Basseville and J.-F. Cardoso, “On entropies, divergences and mean values,” in *Proceedings of the IEEE Workshop on Information Theory*, 1995.
- [15] F. Nielsen and S. Boltz, “The Burbea-Rao and Bhattacharyya centroids,” *Computing Research Repository (CoRR)*, vol. <http://arxiv.org/>, April 2010.
- [16] S. Amari and H. Nagaoka, *Methods of Information Geometry*. A. M. Society, Ed. Oxford University Press, 2000.
- [17] F. Nielsen and V. Garcia, “Statistical exponential families: A digest with flash cards,” 2009, [arXiv.org:0911.4863](http://arxiv.org/0911.4863).
- [18] J.-D. Boissonnat, F. Nielsen, and R. Nock, “Bregman voronoi diagrams,” *Discrete and Computational Geometry*, April 2010. [Online]. Available: <http://dx.doi.org/10.1007/s00454-010-9256-1>
- [19] T. Kailath, “The divergence and Bhattacharyya distance measures in signal selection,” *Communications, IEEE Transactions on [legacy, pre - 1988]*, vol. 15, no. 1, pp. 52–60, 1967.
- [20] T. W. Rauber, T. Braun, and K. Berns, “Probabilistic distance measures of the dirichlet and beta distributions,” *Pattern Recogn.*, vol. 41, no. 2, pp. 637–645, 2008.
- [21] H. Avi-Itzhak and T. Diep, “Arbitrarily tight upper and lower bounds on the bayesian probability of error,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 1, pp. 89–91, 1996.
- [22] D. H. Johnson and S. Sinanovic, “Symmetrizing the Kullback-Leibler distance,” *Technical report*, Mar. 2001.

Frank Nielsen defended his PhD thesis on Adaptive Computational Geometry in 1996 (INRIA/University of Sophia-Antipolis, France), and his accreditation to lead research in 2006. He is a researcher of Sony Computer Science Laboratories Inc., Tokyo (Japan) since 1997, and a professor at École Polytechnique since 2008. His research focuses on computational information geometry with applications to visual computing.